IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 28, 2019, accepted December 24, 2019, date of publication December 30, 2019,
date of current version January 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962974

Robust Head Pose Estimation Using Extreme Gradient Boosting Machine on Stacked Autoencoders Neural Network

MINH THANH VO[1], TRANG NGUYEN[2], AND TUONG LE[3]
[1]Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam
[2]Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City 700000, Vietnam
[3]Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh 700000, Vietnam

Corresponding author: Tuong Le (ct.le@hutech.edu.vn)

ABSTRACT Head pose estimation is an important sign in helping robots and other intelligence machines understand human. It plays a vital role in designing human computer interaction systems because many applications rely on precise results of head pose angles such as human behavior analysis, gaze estimation, 3D head reconstruction etc. This study presents a robust approach for estimating the head pose angles in a single image. More specifically, the proposed system first encodes the global features extracted from Histogram of Oriented Gradients in a multi stacked autoencoders neural network. Based on the hidden nodes in deep layers, Autoencoder has been proposed for feature reduction while maintaining the key information of data. A scalable gradient boosting machine is then employed to train and classify the embedded features. Experiences have evaluated on the Pointing 04 dataset and show that the proposed approach outperforms the state-of-the-art methods with the low head pose angle errors in pitch and yaw as 6.16° and 7.17°, respectively.

INDEX TERMS Head pose estimation, autoencoder, feature reduction, gradient boosting, global features.

I. INTRODUCTION

Recent the past few years, head pose estimation is one of the active problems in facial analysis [1]–[7] that attracts lots of attention from researchers due to its various application in human-computer interaction, face recognition, 3D face modeling, driver monitoring, etc. Specifically, head pose estimation is the task of inferring the orientation of human head from various sources such as: single or sequences 2D images, videos or kinetic sensor with depth information, to name few. A popular problem is to estimate head pose angles from a single 2D image or 2D video and retrieve the computed head pose angles which might include in 3 axes: pitch (x-axis), yaw (y-axis) and roll (z-axis). Although recent researches have been witnessed in many achievements in face recognition, the accuracy of head pose estimation is not satisfied due to many constrains in environment such as illumination condition, facial expression, partially occluded head region, variation head angle changing and other latent variables.

Numerous approaches have been proposed to address with the automatic head pose estimation. These methods could be grouped into several main categories: template-based methods, regression-based methods, deformable model-based methods and manifold based methods. Template based methods [8]–[10] address the head pose estimation as the classification problem where the pose angles are classified to a class of given known pose label. Regression based methods [11]–[14] propose a linear or nonlinear function to map from the training image or extracted features to discrete or continuous pose estimation which often might be affected with noise in environment. Many approaches are derived from deformable model which use a set of parametric models to capture a face model which then is used to match with the testing face [15]–[17]. These methods are often demanded on the facial landmark points to estimate the specific shape which could lead to the constraint to extent and apply in such of low-resolution images. Recently, many approaches have granted success in head pose estimation with promising results by looking a low dimensional representation of head pose angle from high dimensional space. Manifold based methods [18]–[21] assume that discriminative head pose angles lie on the lower dimensional manifold embedded which could be found by the unsupervised or supervised learning. The biggest challenge in

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

manifold learning is to obtain an embedded view which contains the changes due to the pose and ignore other image variation sources such as lightning and noise.

Among those dimensional reduction methods, Autoencoder becomes a promising candidate with many great results achieved in the past few years [22]–[24]. Autoencoder is a combination of two parts: the encoder part for compressing the data in the latent space and the decoder part for reconstructing it to the original information. In dimensionality reduction, Autoencoder tries to find a compact representation of the input through minimizing the error reconstruction between the input and the reconstructed output in neural nodes. Recent researches have shown that Autoencoder could produce meaningful features inside high dimensional data from the embedded space [23], [25], [26].

In recent years, tree-based ensemble methods have got a celebrity status in prediction problem. Instead of using a single model, they combine many weak regression trees with poor performance to improve the prediction accuracy. Another advantage is that they could fix various types of predictor variable while requiring little data preprocessing and handle with nonlinear function. Among with many trees-based ensemble methods and their extensions, Extreme Gradient Boosting (XGB) is a recent proposed method which has been applied in many problems and got many promising results [27] in classification problems.

This paper presents a framework called Stacked Auto Encoder with Extreme Gradient Boosting (SAE-XGB) to estimate the head pose angles from a single 2D image. Face region is first detected and cropped before putting in the preprocessing step. In the preprocessing step, feature information which defines characteristic of pose angle is extracted. In computer vision, global features such as Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), Local Binary Pattern (LBP) etc. include the shape description, contour representation and texture features, which are invariant with noise and illumination change. In this paper, we choose HOG extracted from the face region detected as the input feature vectors. We proposed a multi stacked autoencoder neural network to compress the feature vector from the high dimensional space to a low dimensional representation. The head pose angles are predicted by training the embedded feature vectors with XGB model to improve the accuracy. Overall, our main contribution could be summarized as follows: As far as our best knowledge, this is the first work to propose to use autoencoder neural network for dimensionality reduction in head pose estimation. The model is improved the prediction by employing a scalable gradient boosting system. Finally, the experiments prove that our method outperforms the state-of-the-art head pose estimation methods.

The rest of the paper is organized as follows. Section 2 presents the related work. The proposed framework is presented in Section 3. Experiment will be conducted and discussed in section 4. Finally, section 5 summarizes the results and gives some conclusions.

II. RELATED WORKS

Head pose estimation is an interesting problem which has attracted researches and reviews in a long period of time. Most of popular works mainly use the 2D RGB image data. As mentioned in previous section, those methods could be grouped into several main approaches: template-based methods, regression-based methods, deformable based method and manifold learning methods.

A. HEAD POSE ESTIMATION

Template based methods treat the problem as the classification problem where splitting the training image into several groups with corresponding labels and train a classifier for those groups. Some of well-known works are Yan *et al.* [8] and Li *et al.* [9]. Recent approach from Geng and Xia [10] introduced the concept of soft labelling by considering the neighbor labels around the true pose to alleviate the subjection in collecting training images. Although those approaches could provide sufficient results, they are heavily affected with noise and overfitting.

Regression based methods optimize a mapping from linear or nonlinear functions from feature inputs to continuous or discrete pose value, with possible candidate models such as Support Vector Regressor from work of Ma *et al.* [11], Gaussian Processes Regressors from Ranganathan *et al.* [12]. Another work from Haj *et al.* [13] proposed to correct the misalignment of head location in image by using Partial Least Square (PLS). Recently, Drouard *et al.* [14] proposed to use a mix method from unsupervised learning and regressor in head pose estimation. The regressor model called Partially Latent Mixture of Linear Regressor maps extracted features from high dimensional space to latent space of head pose angles and bounding-box shifts. With many developments in artificial neural networks and convolutional neural network (CNN), various methods proposed to use CNN to compute the head pose estimation, which could be considered as a recent approaches of regression methods. Liu *et al.* [28] proposed to use CNN network to train on a simulated synthetic head image first and then used it to evaluate on the real data. Recent approach from Ranjan *et al.* [29] proposed a complete system to detect even face, landmark localization, pose estimation and gender by using a trained CNN to extract the intermediate features to put it in multi-task learning algorithms. The approach achieved promising results in different problem fields, but it requires a large scale of image to train CNN.

Deformable Based Methods: Another main approach is called deformable based methods where they try to use a statistical model and optimize the model parameters which could describe well with head pose. Commonly method is Active Shape Models (ASMs) [15]. Constraint Local Models [16] tries to minimize the disparity between the 2D feature points and projected 3D points in 2D plane through a rotation process. Another work was from Sun *et al.* [17] when they proposed to use Non-Least Square Model to compute the depth estimation from 3D view for estimating the head pose angles. Those methods are often demanded on landmark points which made them sensitive to several uncontrol factors such as change of identity, occluded part region, or facial expression.

Manifold Learning Methods: In recent years, manifold learning is becoming a promising approach to solve the head pose estimation when many works proposed to use from simple such as PCA, LDA [18] to other advanced embedded methods like Isomap [19], Locally Linear Embedding (LLE) [20]. The idea is root from the assumption that the meaningful pose angles lie in a low-dimensional space. Wang *et al.* [21] proposed a method called Supervised Sparse Manifold Regression to combine the supervised learning and sparse regression into manifold space. Another work from Wang and Song [30] when they improve the head pose accuracy by incorporating the pose angles information into manifold learning space.

The recent advent of depth cameras enabled a new approach for head pose estimation using depth-based information. Depth data helps to improve the weakness in using RGB image only such as noise, illumination change, expression. For an instance, Seemann *et al.* [31] proposed to incorporate the depth information at head region with color histogram to train a neural network. In a different approach, Venturelli *et al.* [32] used Siamese architecture in improving a deep neural network training in a Kinect head pose data. Recently, Zhang *et al.* [33] used a multi-stream multitask neural network to estimate head pose angle in RGB-D videos, which is 2D videos containing with depth information. In many situations, depth is considered as an addition information added with photometric data and cannot be used alone.

B. AUTOENCODER FOR DIMENSIONAL REDUCTION

Autoencoder is one of advanced neural networks that have many applications, one of these is dimensionality reduction. Wang *et al.* [24] took a quantitatively survey to compare the ability to reduce dimensionality of autoencoder and others state-of-the-art dimensionality reduction methods. The work has shown that the autoencoder can indeed learn meaningful somethings in the latent space. Jiang *et al.* [34] proposed to use supervised information to further guide the autoencoder in finding the latent space. Another work [35] combined the hand craft feature extracted in features and latent features extracted by autoencoder to get fast image retrieval in domains. Although this network has been used to solve the dimensionality reduction in various problem [36]–[38], there is no recent works in head pose estimation and related problems.

C. GRADIENT BOOSTING MODELS

Gradient Boosting is an important tool in the field of machine learning, providing with great achievements in performance on classification, regression and ranking tasks. Among all different gradient boosting algorithms, gradient tree boosting [39] is a highly effective and widely used method. It has been shown to achieve the state-of-the-art accuracy on many classification benchmarks [40]. However, to the best of our knowledge, research on gradient tree-based boosting in head pose estimation has not been fully documented to date. Driven by the successful application of methods in various problem field such as travel time prediction [41], predicting symptom severity [42], tracking algorithm [43] etc., we proposed to use a recent advanced tree-based boosting called Extreme Gradient Boosting (XGB) [27] to predict the head pose angles encoded in the embedded space generated by Autoencoder.

III. THE PROPOSED FRAMEWORK

In this section, the proposed method called SAE-XGB is described in detail. The overview framework is given in Figure 1. This model estimates the pitch head pose angles and the yaw head pose angles from a 2D image. Given single query image, the face region in image is detected using Single Shot Multibox Detector [53] and cropped to size 150 × 150 pixels. After that, the Histogram of Oriented Gradient (HOG) features in image are extracted. In this paper, we use both unsigned and signed histogram bin. In addition, additional energy functions are used to provide the robust information for image. In specific:

- Image is divided in 9×9 blocks
- Use 9 unsigned bins, 18 signed bins and 4 energy features to compute the feature vectors
- In total, that gives us a one-dimensional vector $v \in \mathbb{R}^{1 \times D}$, with $D = 9 \times 9 \times (9 + 18 + 4) = 2511$.

To further reduce dimensionality, multi fully connected hidden layers with different hidden units are stacked for forming a Stacked Autoencoder (SAE), as shown in Figure 2. Autoencoder is an unsupervised learning neural network having structure very similar to the multiplayer perceptron (MLP) network which often have an input layer, an output layer. It is noticed that the output layer would have the same node as the input layer to reconstruct its own input. The autoencoder has two parts: encoder part and decoder part. In this study, we reduce them through two fully connected layer having hidden units 512 and 256, respectively. Each fully connected layer is followed by Rectifier layer (ReLU), a nonlinear layer help model generalizes with variety of data. Next, a Dropout layer is put to randomly drop 20% probability of hidden units in training process for preventing the overfitting problem. The SAE network in trained by Adam optimized method in 10 epochs with batch size 30 at learning rate 0.001. The reduced features extracted from the second fully connected layer would have the length of 256.

Next, those features vectors are then trained a classifier namely XGB to classify two classes pitch and yaw angle respectively in 100 iterations at learning rate 0.05. XGB is a supervised learning model, proposed by Chen and Guestrin [27]. Its core uses Gradient Boosting Machine model which combines weak ''learner'' into strong classifier in an iterative process.

In the testing phase, face region is detected and extracted from the single query image. In the following step,

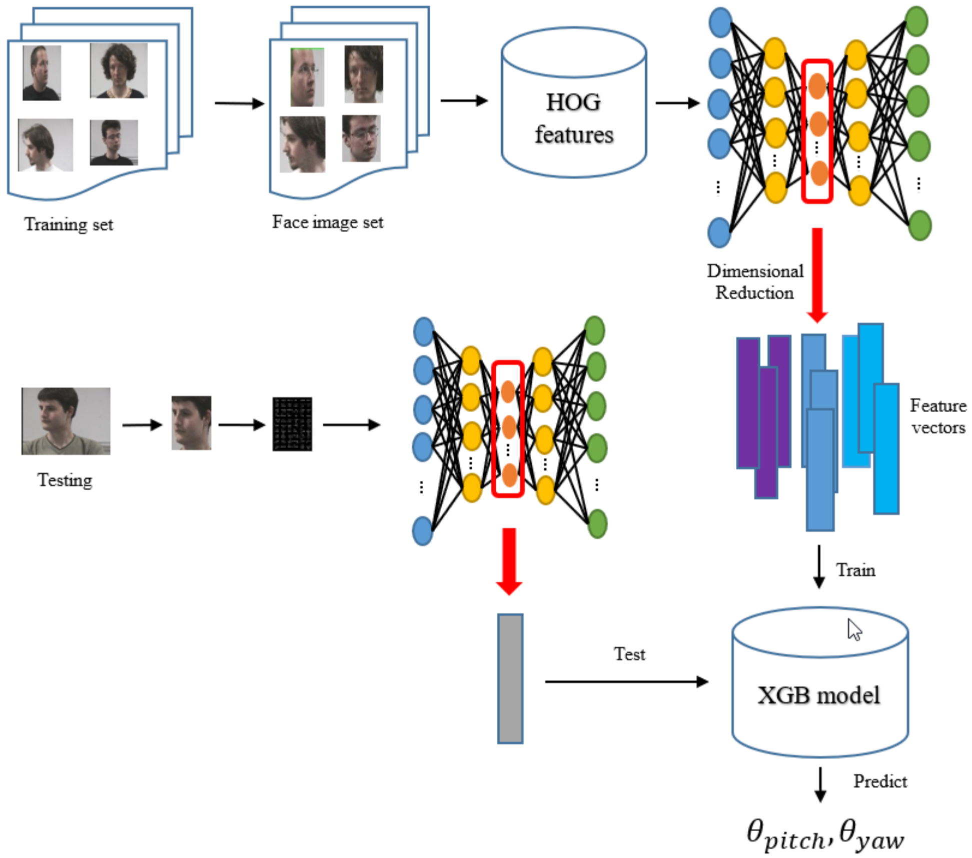


FIGURE 1. The proposed SAE-XGB model for head pose estimation.

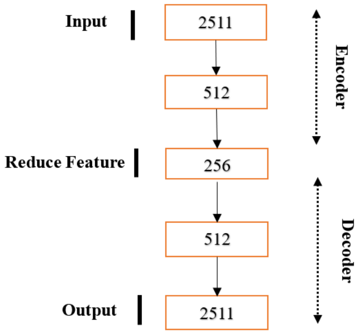


FIGURE 2. Overview on SAE network.

HOG features are extracted and flatten into vector where it is put to trained stack autoencoder model (SAE). The encoder of SAE helps to encode the vector into the more compact form with length of 256. The predicted head pose angle, specific to pitch and yaw rotation are estimated by putting the reduced feature vector through the trained XGB model.

IV. EXPERIMENT STUDY

A. POINTING 04 DATASET

In this section, we describe the description about Pointing 04 dataset [44]. It consists of 15 sets of people which have various poses. There are 13 yaw angles $\{-90, -75, -60, -45, -30, -15, 0, 15, 30, 45, 60, 75, 90\}$ and 9 pitch angles $\{-90, -60, -30, -15, 0, 15, 30, 60, 90\}$ degrees. The pose angles are recorded by asking the participant to look at 93 markers which is corresponding to a specific pose in two times. There are 93 poses available for each person, which makes total $93 \times 15 \times 2 = 2790$ images. The dataset is manually annotated with a face bounding box. Figure 3 presents some samples from the Pointing 04 dataset with their pitch and yaw head pose angles, respectively.

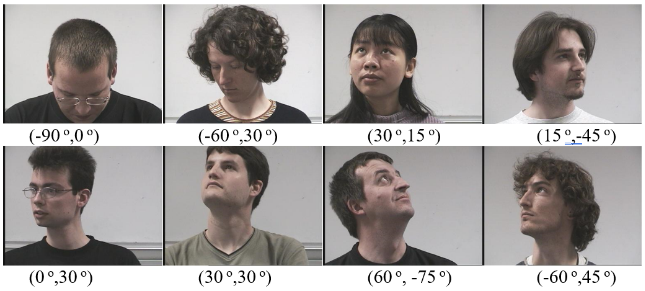


FIGURE 3. Sample images with corresponding pitch and yaw head pose angles in the Pointing 04 dataset, respectively.

1) DATA TRAINING

We perform the cross-validation on the dataset where it is divided to six folds. Five of six subsets are used to train the SAE-XGB model while the last subset is used for testing. This procedure is repeated 6 times. The SAE-XGB model contains two training components: the stack autoencoder (SAE) component and the XGB component. In each repetition,

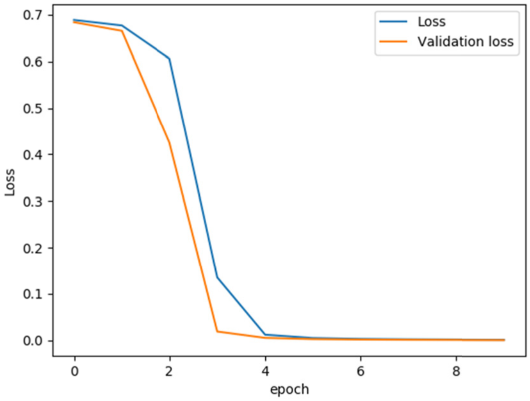


FIGURE 4. Training loss and validation loss when training SAE component.

we use five of six subsets to train SAE with 80% of dataset for training and 20% for validation. Figure 4 present the loss when training SAE component of one running repetition time. As we can see that, the training loss and validation loss decrease dramatically just after a few epochs and become stable after 4 epochs, which has proven that the SAE model could quickly train and find the generalization of dataset.

After training, the encoder of SAE model is used to transform the dataset of all six folds from original feature space to a latent feature space of 256. It is notice that dataset in the last fold is kept separately and is not used in the training procedure. We found that the new latent feature input transformed SAE model could present a more meaningful meaning. Figure 5 shows the feature reduction applied on the dataset of six folds using some popular manifold learning methods such as principal component analysis (PCA), linear discriminant analysis (LDA) compared with feature reduction using our stack autoencoder model. It is clearly to see that the embedded input transformed from SAE model could group data points of same classes to the same regions much better than using other feature reduction methods. This has proven that using SAE, we could represent data in a new compact feature space with much more meaningful than the original data. However, when we reduce the input from the original feature space to a lower feature space, the information of data will be lost. In this case, the three-dimensional space might be good for visualization, but the information of dataset will be affected significantly. We balance between the lower feature space of data and the lost information by choosing a higher feature space of 256.

The feature input transformed from SAE model will be trained using XGB component. Again, five of six folds will be used to train XGB model while the last fold is used for testing. The mean absolute error (MAE) and the accuracy of the predicted head pose angle and the ground truth pose will be calculated and reported. We trained XGB with 100 epochs and measure the logarithmic loss in the training process, which is defined as followed:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i logp_i + (1-y_i)\log(1-p_i)\right], \qquad (1)$$

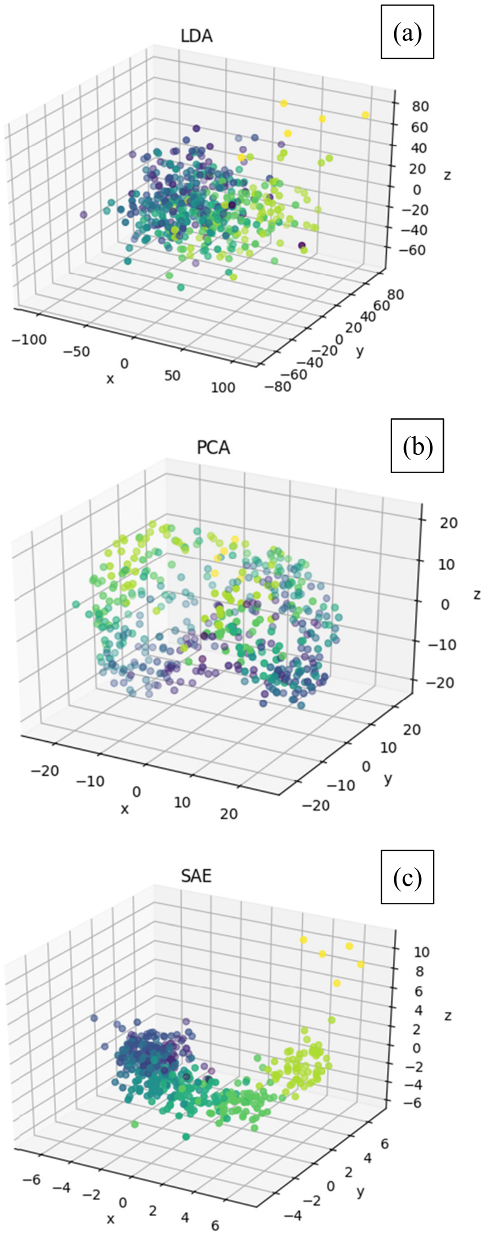


FIGURE 5. Feature reduction of input dataset in three-dimensional space using various methods. (a) LDA method (b) PCA method (c) SAE model. The embedded feature points got from SAE model could be grouped in same regions much better than using other feature reduction methods such as PCA, LDA.

where N is the number of training samples, y_i is the outcome of the i-th instance and p_i would be the probability of the i-th instance having the value y_i. Figure 6 presents the log loss of XGB model training in one validation of the cross-validation. The figure shows that the training log loss decreases quickly and becomes stable under 0.25 after 40 epochs.

B. EXPERIMENTAL RESULTS

In the first experiment, we report the accuracy of the proposed model in cross-validation. Table 1 provides detail classification accuracy of the predicted pitch angles and yaw angles

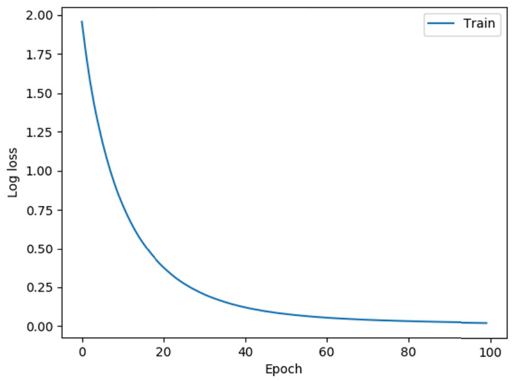


FIGURE 6. The log loss of XGB training procedure in one-fold.

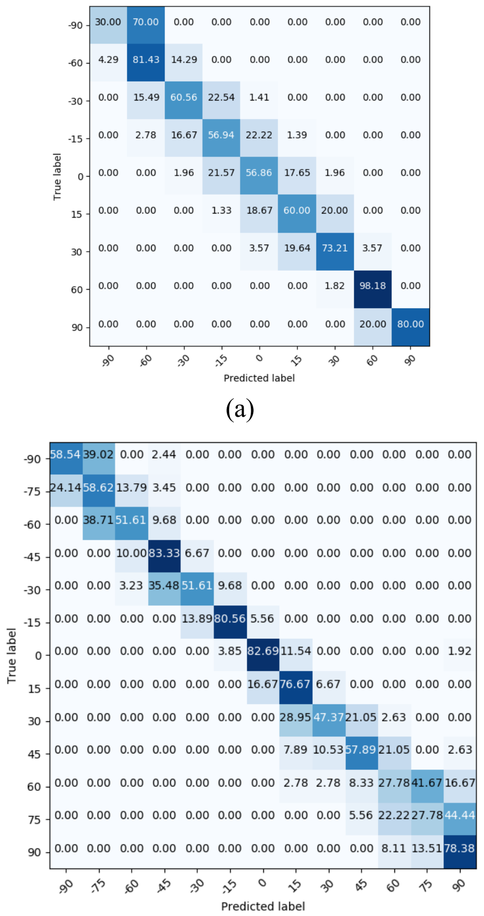
TABLE 1. The accuracy of pitch and yaw head pose angles in each fold.

Fold	Accuracy (%)	
	Pitch	Yaw
Fold 1	67.31	61.51
Fold 2	68.82	59.14
Fold 3	69.89	55.27
Fold 4	70.54	51.18
Fold 5	70.54	60.65
Fold 6	66.88	55.70

in each fold. As could be seen in table, the pitch head pose prediction performs slightly better than the yaw angle.

In greater detail, the confusion matrices (in %) computed in one-fold of the proposed methods on pitch and yaw respectively are reported in Figure 7. As been shown clearly, most incorrect predictions are adjacent to the proper ground truth angle. This suggests that although the model predicts the wrong head pose angles, the differences between the predicted values and ground truth values are small and acceptable in practical problems when the constrains of head pose angles are in differences of 5 degrees. In additional, the errors tend to become increasing at the larger head pose angles. It is noticed that the model performs badly at the −90° pitch angles than other since there are much fewer training examples for this angle.

In the next following experiment, we do the leave-one-out testing where using 14 subjects for training while keeping the last subject separately in the testing (unknown testing). The validation process is taken and repeated the testing subject randomly 6 times. Table 2 represents the mean of accuracy ± standard deviation of the validation and several state-of-the-art methods reported on head pose estimation problem including neural network based method [45], [46], the method using partial least square approach [47] and the approach using a tensor model to evaluate head pose angles based on the location of nose-tip [48] and even the start-of-the method using CNNs model [52]. The best mean performance is highlighted by bold face. According to experimental results, our proposed method achieves better than all other compared methods on evaluation measure.



(b)

FIGURE 7. The confusion matrices (in %) of proposed method on (a) pitch angles and (b) yaw angles.

TABLE 2. Head pose estimation classification accuracy of various methods on the Pointing 04 dataset.

Method	Accuracy (%)	
	Pitch	Yaw
SAE-XGB	**68.99±1.46**	**57.24±3.56**
Patacchiola (CNNs) [52]	61.4	62.33
Stiefelhagen [45]	66.3	52.0
Haj (Linear PLS) [13]	58.70	45.57
Human Performance [46]	59.0	40.7
Gourier (Associative Memories) [46]	43.9	50.0
Tu (High-order SVD) [48]	54.84	49.25
Tu (PCA) [48]	57.99	55.20
Tu (LEA) [48]	50.61	45.16

When dealing with head pose estimation, the problem is often evaluated based on the regression measures between the predicted pose and ground truth pose values. The popular used measure metric is mean absolute error (MAE), which is defined as follow.

$$MAE = \frac{1}{N} \sum\nolimits_{i=1}^{N} |\bar{x}_i - x_i| , \tag{2}$$

with $\bar{x}_i$ is the predicted head pose and x_i is the ground truth head pose angle.

TABLE 3. The MAE for pitch and yaw angles of various methods performed in Pointing 04 dataset.

Method	Head Pose Estimation (Degrees)	
	Pitch	Yaw
SAE-XGB	**6.16±0.31**	**7.17±0.66**
SAE-XGB(SIFT)	8.17	9.45
SAE-XGB(LBP)	10.11	11.47
Patacchiola (CNNs) [52]	10.57	7.74
Stiefelhagen [45]	9.7	9.5
Haj (Linear PLS) [13]	10.52	11.29
Human Performance [46]	9.4	11.8
Gourier (Associative Memories) [46]	15.9	10.1
Tu (High-order SVD) [48]	17.97	12.9
Tu (PCA) [48]	14.98	14.11
Tu (LEA) [48]	17.44	15.88
Drouard [14]	8.47±10.35	7.93±7.9
GPR [50]	11.94±10.19	15.04±12.24
PLS [47]	12.25±9.73	13.38±10.8
SVR [49]	11.25±9.42	12.82±10.99

Then, the proposed method is compared with some previous mentioned above methods. We additionally benchmarked the following support vector regression [49], Gaussian regression model [14], [50] as they are widely popular and used in classification and regression problem. As be shown clearly in Table 3, SAE-XGB model yields the lowest errors for both pitch and yaw in compared with all other methods. Clearly, it shows that features vectors reduced from stack autoencoder network contribute to the discrimination of head pose angles information in the vector space. In additional, the XGB model has proven the stability in predicting head pose angles under various head poses.

In addition, we provide an extensive comparison between the feature we use HOG with other common visual descriptors including Scale-invariant Feature Transform (SIFT) and Local Binary Pattern (LBT) denoted by SAE-XGB(SIFT) and SAE-XGB(LBT), respectively. While HOG computes the number of occurrences of gradient orientation in the crop face image, SIFT finds key points of in images by applying Difference of Gaussian and finding local extrema over scale and space. LBP, in the other hand, constructs histogram formed a binary coding vector surrounding pixel center. Looking at the Table 3, we could see that using HOG in SAE-XGB achieve a better performance than other methods in capture the description of head pose information.

V. CONCLUSION

This paper presents a robust framework called SAE-XGB to estimate head pose angles in various poses. Face images are collected and then extracted the prominent information to create feature vectors. The framework proposed a stack autoencoder neural network training on those features vectors for learning a dimensional reduction way to reduce features vectors from high dimensional to lower representation. The reduced features are trained on a classifier which use XGB model to predict head pose angles. Experiments have shown the effectiveness of SAE-XGB method. The proposed method is compared with many state-of-the-art head pose estimation algorithms on the Pointing 04 dataset and yields significantly better results not only in accuracy but the MAE scores also. Specifically, SAE-XGB achieves in accuracy 68.99% for pitch angles prediction and 57.24% for yaw angles prediction. In additional, it gets the MAE with 6.16o for pitch and 7.17o for yaw angles, respectively. In future works we will use the results of this study, head pose estimation, to improve the accuracy of gaze tracking module embedded in human computer interaction system.

REFERENCES

[1] D. Cui, G.-B. Huang, and T. Liu, "ELM based smile detection using distance vector," *Pattern Recognit.*, vol. 79, pp. 356–369, Jul. 2018.

[2] B. Alsalibi, I. Venkat, and M. A. Al-Betar, "A membrane-inspired bat algorithm to recognize faces in unconstrained scenarios," *Eng. Appl. Artif. Intell.*, vol. 64, pp. 242–260, Sep. 2017.

[3] W. Hariri, H. Tabia, N. Farah, A. Benouareth, and D. Declercq, "3D facial expression recognition using kernel methods on Riemannian manifold," *Eng. Appl. Artif. Intell.*, vol. 64, pp. 25–32, Sep. 2017.

[4] J. Gou, L. Wang, Z. Yi, J. Lv, Q. Mao, and Y.-H. Yuan, "A new discriminative collaborative neighbor representation method for robust face recognition," *IEEE Access*, vol. 6, pp. 74713–74727, 2018.

[5] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," *IEEE Access*, vol. 7, pp. 23012–23026, 2019.

[6] T. Vo, T. Nguyen, and C. T. Le, "A hybrid framework for smile detection in class imbalance scenarios," *Neural Comput. Appl.*, vol. 31, no. 12, pp. 8583–8592, Dec. 2019.

[7] T. Vo, T. Nguyen, and C. Le, "Race recognition using deep convolutional neural networks," *Symmetry*, vol. 10, no. 11, p. 564, Nov. 2018.

[8] S. Yan, H. Wang, J. Tu, X. Tang, and T. Huang, "Mode-kn factor analysis for image ensembles," *IEEE Trans. Image Process.*, vol. 18, no. 3, pp. 670–676, Mar. 2009.

[9] S. Li, Q. Fu, L. Gu, B. Scholkopf, Y. Cheng, and H. Zhang, "Kernel machine based learning for multi-view face detection and pose estimation," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2002, pp. 674–679.

[10] X. Geng and Y. Xia, "Head pose estimation based on multivariate label distribution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1837–1842.

[11] Y. Ma, Y. Konishi, K. Kinoshita, S. Lao, and M. Kawade, "Sparse Bayesian regression for head pose estimation," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 507–510,

[12] A. Ranganathan, M.-H. Yang, and J. Ho, "Online sparse Gaussian process regression and its applications," *IEEE Trans. Image Process.*, vol. 20, no. 2, pp. 391–404, Feb. 2011.

[13] M. A. Haj, J. Gonzalez, and L. S. Davis, "On partial least squares in head pose estimation: How to simultaneously deal with misalignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2602–2609.

[14] V. Drouard, R. Horaud, A. Deleforge, S. Ba, and G. Evangelidis, "Robust head–pose estimation based on partially–latent mixture of linear regressions," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1428–1440, Mar. 2017.

[15] T. Cootes, G. Wheeler, K. Walker, and C. Taylor, "View-based active appearance models," *Image Vis. Comput.*, vol. 20, nos. 9–10, pp. 657–664, Aug. 2002.

[16] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean–shift," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 200–215, Jan. 2011.

[17] Z.-L. Sun, K.-M. Lam, and Q.-W. Gao, "Depth estimation of face images using the nonlinear least–squares model," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 17–30, Jan. 2013.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). Berlin, Germany: Springer-Verlag, 2006.

[19] J. B. Tenenbaum, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.

[20] S. T. Roweis, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[21] Q. Wang, Y. Wu, Y. Shen, Y. Liu, and Y. Lei, "Supervised sparse manifold regression for head pose estimation in 3D space," *Signal Process.*, vol. 112, pp. 34–42, Jul. 2015.

[22] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[23] W. Wang, Y. Huang, Y. Wang, and L. Wang, "Generalized autoencoder: A neural network framework for dimensionality reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014.

[24] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, Apr. 2016.

[25] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008.

[26] S. Rifai, P. Vincent, X. Müller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 833–840.

[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2016, pp. 785–794

[28] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei, "3D head pose estimation with convolutional neural network trained on synthetic images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1289–1293.

[29] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi–task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[30] C. Wang and X. Song, "Robust head pose estimation via supervised manifold learning," *Neural Netw.*, vol. 53, pp. 15–25, May 2014.

[31] E. Seemann, K. Nickel, and R. Stiefelhagen, "Head pose estimation using stereo vision for human-robot interaction," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Jun. 2004, pp. 626–631.

[32] M. Venturelli, G. Borghi, R. Vezzani, and R. Cucchiara, "From depth data to head pose estimation: A siamese approach," in *Proc. 12th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2017.

[33] G. Zhang, J. Liu, H. Li, Y. Q. Chen, and L. S. Davis, "Joint human detection and head pose estimation via multistream networks for RGB–D videos," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1666–1670, Nov. 2017.

[34] X. Jiang, X. Song, J. Gao, Z. Cai, and D. Zhang, "Nonparametrically guided autoencoder with Laplace approximation for dimensionality reduction," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016.

[35] S. Petscharnig, M. Lux, and S. Chatzichristofis, "Dimensionality reduction for image features using deep learning and autoencoders," in *Proc. 15th Int. Workshop Content-Based Multimedia Indexing (CBMI)*, no. 23, 2017.

[36] N. M. Peleato, R. L. Legge, and R. C. Andrews, "Neural networks for dimensionality reduction of fluorescence spectra and prediction of drinking water disinfection by-products," *Water Res.*, vol. 136, pp. 84–94, Jun. 2018.

[37] S. A. Taghanaki, J. Kawahara, B. Miles, and G. Hamarneh, "Pareto-optimal multi-objective dimensionality reduction deep auto-encoder for mammography classification," *Comput. Methods Programs Biomed.*, vol. 145, pp. 85–93, Jul. 2017.

[38] Y. Nozaki and T. Nakamoto, "Itakura–Saito distance based autoencoder for dimensionality reduction of mass spectra," *Chemometrics Intell. Lab. Syst.*, vol. 167, pp. 63–68, Aug. 2017.

[39] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[40] P. Li, "Robust logitboost and adaptive base class (ABC) logitboost," in *Proc. 26th Conf. Annu. Conf. Uncertainty Artif. Intell. (UAI)*, 2010, pp. 302–311.

[41] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 308–324, Sep. 2015.

[42] Y. Liu, Y. Gu, J. C. Nguyen, H. Li, J. Zhang, Y. Gao, and Y. Huang, "Symptom severity classification with gradient tree boosting," *J. Biomed. Informat.*, vol. 75, pp. S105–S111, Nov. 2017.

[43] J. Son, I. Jung, K. Park, and B. Han, "Tracking-by-segmentation with online gradient boosting decision tree," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015.

[44] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial features," in *Proc. Int. Workshop Vis. Observ. Deictic Gestures*, 2004.

[45] R. Stiefelhagen, "Estimating head pose with neural networks-results on the pointing04 ICPR workshop evaluation data," in *Proc. Pointing Workshop, Vis. Observ. Deictic Gestures*, Cambridge, U.K., 2004.

[46] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley, "Head pose estimation on low resolution images," in *Proc. 1st Int. Workshop Classification Events, Activities Relationships*, 2006, pp. 270–280.

[47] H. Abdi, "Partial least square (PLS) regression," in *Encyclopedia for Research Methods for Social Sciences*. Thousand Oaks, CA, USA: SAGE, 2003, pp. 792–795.

[48] J. Tu, Y. Fu, Y. Hu, and T. Huang, "Evaluation of head pose estimation for studio data," in *Proc. 1st Int. Workshop Classification Events, Activities Relationships*, 2006, pp. 281–290.

[49] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

[50] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2005.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015.

[52] M. Patacchiola and A. Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," *Pattern Recognit.*, vol. 71, pp. 132–143, Nov. 2017.

[53] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 1, 2016, pp. 31–37.

MINH THANH VO received the B.S. degree from the University of Science, Ho Chi Minh City, Vietnam, in 2012, and the M.S. degree from Sejong University, South Korea, in 2019. He joined the Institute of Research and Development, Duy Tan University, Vietnam, in 2019, as a Researcher. His interests are machine learning, deep learning, imbalanced data problem, data science, and computer vision.

TRANG NGUYEN received the B.S. degree from Hung Vuong University, Vietnam, in 2010, and the M.S. degree from the University of Information Technology, Vietnam, in 2014, where she is currently pursuing the Ph.D. degree. She is also a Lecturer with the Faculty of Information Technology, Ho Chi Minh City Open University, Vietnam. Her research interests are natural language processing, artificial intelligence, social network analysis, cybernetic and systems, collaborative, ontology building, semantic search, and deep learning.

TUONG LE is currently a Lecturer and a Researcher with the Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh, Vietnam. His research interests are machine learning, imbalanced learning, deep learning, business intelligence, data analysis, data mining, and pattern mining. He has published more than 25 articles in prestigious journals, such as *Information Sciences*, *Expert Systems with Applications*, IEEE Access, and *Engineering Applications of Artificial Intelligence*. He served as a Reviewer for several journals like IEEE Transactions on Cybernetics, IEEE Access, *Applied Soft Computing*, *Neural Computing and Applications*, *Applied Intelligence*, *PLOS ONE*, and *Engineering Applications of Artificial Intelligence*.